We are IntechOpen,
the world's leading publisher of
Open Access books
Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Spectral Analysis of Exons in DNA Signals

Noor Zaman, Ahmed Muneer and
Fausto Pedro García Márquez

Additional information is available at the end of the chapter

1. Introduction

DNA is found in blood cells carrying nucleus. The DNA is isolated from blood through a series of different procedures including heat shock, thermal change and applications of different chemicals etc. DNA sequence contains chromosomes which further contains genes over them. The genes have regions which could translate to protein and the regions which don't perform any contribution in protein production. Both kinds of regions are made-up of nucleotides characterized as Adenine, Thymine, Cytosine and Guanine. The order of these nucleotides determines the traits, habits and livings of all species. Since with the exponential growth of biological data, there is an enormous amount of such data that needs to be translated to protein. A successful translation would result in knowing important information about species.

Comparative analysis of computational techniques employed over genetic datasets has given very interesting results. We are able to identify species from each other on the behalf of DNA properties. A true correct conversion takes to fruitful results. Literature has shown that direct comparative analysis is not as useful as approximate estimation. So far, there is no compact solution available that could outperform for a robust translation from DNA to RNA.

It is a common phenomenon that nucleotide sequences in DNA perform a period three property [3, 11] due to codon composition and structure in the strand. This fundamental characteristic can be exploited to predict the codon regions that help in determination of RNA sequences in DNA. This finding is of immense importance as cell growth and function is determined by the type of protein the cell produces and helps in drug design and revealing genetic disorders as a result of mutation in structure of nucleotide bases (order in which they appear over chain). Many approaches have been proposed in literature that addresses this open optimization problem in computational biology.

Discrete Fourier Transforms [6, 7, and 8] normally result in spectral leakage that doesn't preview the optimal power spectral density estimation. On the other hand, the Short Time Fourier Transforms [2, 4] minimize the leakage but are considered useful when we desire to have the frequency contents with location information. It can plot the components for time, amplitude and frequency of a genetic signal.

Digital Filters [5, 7, and 13] present the spectral contents of signal around the periodicity property of coding regions but don't specify the frequency time relationship with amplitude.

Dosay-Akbulut [14] emphasized the classification of introns in two groups based on RNA secondary structure and self splicing ability in variant species using PCR.

A. Parent et al., [15] describe the importance of coordination between transcription and RNA processing that carboxy-terminal domain of RNA polymerase II acts as a common link in both.

Al Wadi et al. [16] used wavelet transforms for forecasting volatility in experimental results. M. Hashemi et al. [17] provided Identification of *Escherichia coli* O157:H7 Isolated from Cattle Carcasses in Mashhad Abattoir by Multiplex PCR.

A. Ali et al. [18] have presented a Histopathological Study for development of a model for Tumor Lung Cancer Assessing Anti-neoplastic Effect of PMF in Rodents.

J. Singh et al. [19] proposed a technique for Prediction of in vitro Drug Release Mechanisms from Extended Release Matrix Tablets.

2. Proposed approach

The proposed approach consists of a series of components that analyze the DNA signal and enhances the prediction accuracy of genic regions over DNA sequence. The major steps of proposed approach are,

• Conversion of target DNA stretch to a digital pattern employing an indicator sequence

• Decomposition of signal using wavelet transforms

• Calculations of approximate coefficients of signal at level three

• Calculations of detail coefficients of signal

• Density estimation of signal

• Signal analysis for denosing

• Depiction of original and synthesized signal at level three

• Histogram estimations of signal

• Signal extension to a desired length

- Shannon entropy calculation of signal

- Magnitude and power estimation of signal

- Calculation of discrimination measure for PSD analysis

- Exon and intron boundaries' estimation

As an elaboration, the DNA sequence is passed through a filter that transforms it into a digital pattern. This phase is accomplished employing an indicator sequence with the following weights for nucleotides,

Adenine (A) = X (A) = 0.260
Thymine (T) = X (T) = 0.375
Guanine (G) = X (G) = 0.125
Cytosine (C) = X (C) = 0.370

The corresponding transform becomes

$$X_{\text{IndSeq}}[k] = \sum_{n=1}^{N} x_{\text{IndSeq}}[n]e^{-j2\pi kn/N}$$
$$k = 1,\ 2,...,N$$

(1)

Indicator sequence

The signal is decomposed employing the wavelet transforms of order three at level three

$$
\begin{aligned}
y(t) &= A_1(t) + D_1(t) \\
&= \sum_k cA_2(k)\phi_{j-2,k}(t) + \sum_k cD_2(k)w_{j-2,k}(t) + \sum_k cD_1(k)w_{j-1,k}(t) \\
&= A_2(t) + D_2(t) + D_1(t) \\
&= A_3(t) + D_3(t) + D_2(t) + D_1(t) \\
&= \sum_k cA_3(k)\phi_{j-3,k}(t) + \sum_k cD_3(k)w_{j-3,k}(t) + \sum_k cD_2(k)w_{j-2,k}(t) + \sum_k cD_1(k)w_{j-1,k}(t)
\end{aligned}
$$

(2)

3rd order wavelet decomposition

The wavelet decomposition passes the signal into a series of low and high pass filters that decompose and synthesize the signal for reducing flicker noise (pink noise).

The signal is then convoluted with a window function (Kaiser Window) defined below,

$$w(n) = \begin{cases} I_0\left(\beta\left(1-\left((n-\alpha)/\alpha\right)^2\right)^{1/2}\right)/I_0(\beta) & 0 \leq n \leq M-1 \\ 0 & otherwise \end{cases}$$

(3)

Kaiser window of length 351 bp

Each section of the signal is traversed for calculation of absolute and power values. Each segment is plotted over the power spectral graph keeping the period three property maintained at each step. The exon boundaries appear as sharp peaks. The final discrimination measure depicts the degree of relevance in exon and introns.

3. Results and discussions

A specimen gene pattern S.cerevisiae chromosome III (AF099922) has been taken for experiments over proposed approach. The gene is passed through the series of steps defined,

At processing stage, the dataset is passed through two kinds of filters. First filter refines the data and outputs a data file that purely contains nucleotide characters. Second filter operates on output file obtained from first filter application and generates a file that contains numeric data. This data is fed into central engine for further processing.

Figure 1 Shows dataset that contains nucleotide characters and some other characters. This is first necessary step because this input when fed into our engine will badly degrade the performance and brings false results.

```
 gttaatgtag cttaataaca aagcaaagca ctgaaaatgc ttagatg
gtaaaat tacacatgca121 aacctccata gaccggtgta aaatccc
gacacct tgcctagcca cacccccacg ggactcagca 241 gtgata
01 atttcgtgcc agccaccgcg gtcatacgat taacccaaac taat
atccaac ttatatgtga aaattcattg  421 ttaggaccta aactc
ttagata ccccactatg cttagccata aacctaaata attaaattta
ggcggta601 ctttatatcc atctagagga gcctgttcta taatcga
cagcaaa ccctaaaaag gtattaaagt 721 aagcaaaaga atcaaa
tcttata aaagaacatt actataccct ttatgaaact aaaggactaa
agtacgc 901 acacaccgcc cgtcaccctc ctcaaattaa attaaa
cgtaaca aggtaagcat actggaaagt gtgcttggaa1021 taatca
gacactc tgaactaatc ctagccctag ccctacacaa atataattat
tcgtaca tctaggagct1201 atagaactag taccgcaagg gaaaga
cttttgc ataatgaact aactagaaaa cttctaacta aaagaattac
ctcgtct1381 atgtggcaaa atagtgagaa gatttttagg tagagg
```

Figure 1. Preprocessed dataset

Figure 2 represents a data glimpse that contains pure nucleotide characters.

```
gttaatgtagcttaataacaaagcaaagcactgaaaatgcttagatgga
atgcaaacctccatagaccggtgtaaaatcccttaaacatttgcttaaa
ggactcagcagtgataaatattaagcaataaacgaaagtttgactaagt
attttcggcgtaaaacgtgtcaactataaataaataaatagaattaaaa
tataatacacgacagctaagacccaaactgggattagatacccactat
ttaaaactcaaaggacttggcggtactttatatccatctagaggagcct
cttcagcaaacccctaaaaaggtatttaagtaagcaaaagaatcaaacat
acattactatatcctttatgaaactaaaggactaaggaggatttagtag
ctcaaattaaattaaacttaacataattaatttctagacatccgtttat
ttaatattaaagcatctggcctacacccagaagatttcatgaccaatgg
atttatcctactaaaagtattggagaaagaaattcgtacatctaggagc
ccttgtacctttgcataatgaactaactagaaaacttctaactaaaag
atgtggcaaaatagtgagaagattttaggtagaggtgaaaagcctaac
aacaaatcaaaaagtaagtttaaattatagccaaaagagggacagctc
aaaagcagccaccaacaaagaaagcgttcaagctcaacataaaatttca
caacactgttagtatgagtaacaagaattctaattctccaggcatacac
```

Figure 2. Refined dataset

The EIIP indicator sequence transforms the nucleotides in numeric values as per its defini-
tion. A part of signal is described in Figure 3 below using EIIP indicator sequence as,

```
0.1335 0.1335 0.0806 0.1260 0.1260 0.1335 0.1335 0.1340 0.1260 0.1260 0.1335
0.1340 0.1260 0.1335 0.0806 0.1340 0.1335 0.1335 0.1335 0.1335 0.1335 0.1335
0.1335 0.1260 0.1260 0.1260 0.1260 0.1260 0.0806 0.1260 0.0806 0.1340 0.1260
0.1260 0.1260 0.1260 0.1335 0.1335 0.1335 0.1335 0.1335 0.1335 0.1340 0.1260
0.0806 0.0806 0.0806 0.1260 0.1260 0.1260 0.1335 0.1340 0.1340 0.0806 0.1335
0.0806 0.1340 0.1335 0.1340 0.1260 0.1260 0.1335 0.1335 0.1335 0.1335 0.0806
0.1260 0.1260 0.1340 0.1335 0.1335 0.1260 0.0806 0.1335 0.0806 0.1340 0.1340
0.1335 0.1335 0.0806 0.1340 0.1335 0.1340 0.1340 0.1260 0.1340 0.1340 0.0806
0.1340 0.1335 0.1260 0.1260 0.1260 0.1335 0.1260 0.1335 0.1335 0.1335 0.1340
0.0806 0.0806 0.1260 0.1335 0.1335 0.1335 0.1340 0.1260 0.1260 0.1260 0.1335
0.1260 0.1260 0.1340 0.1260 0.1335 0.0806 0.1260 0.1260 0.0806 0.1335 0.1340
0.0806 0.1340 0.1335 0.1340 0.1335 0.0806 0.0806 0.1260 0.0806 0.1335 0.1335
0.1340 0.0806 0.1260 0.1260 0.1260 0.1260 0.1340 0.0806 0.1260 0.1260 0.1260
0.1340 0.1260 0.1335 0.1335 0.1335 0.1335 0.1335 0.1335 0.1340 0.0806 0.1340
0.1260 0.1260 0.1260 0.1260 0.1260 0.1260 0.1260 0.1260 0.1340 0.0806 0.1340
0.1260 0.1260 0.1340 0.0806 0.1340 0.0806 0.1260 0.1260 0.1260 0.1260 0.1260
0.0806 0.0806 0.1260 0.0806 0.1335 0.1340 0.0806 0.1340 0.1340 0.1260 0.1340
0.0806 0.1340 0.1260 0.1335 0.1335 0.1335 0.1260 0.1335 0.1335 0.1260 0.0806
```

Figure 3. Numeric translation of gene F56F11.5 (AF099922)

The binary indicator sequence is formed by replacing the individual nucleotides with values
either 0 or 1. 1 stands for presence and 0 for absence of a particular nucleotide in specified
location in DNA signal,

Figure 4 describes the glimpse of binary indicator sequence which is the one of four parts of translation of gene file. Only 1's and 0's are visible in this sequence.

```
0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0
1 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 0 1 1
0 0 0 0 0 0 0 1 0 1 1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 0 0 1 0
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
1 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 1 0 1 1 0 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1
0 1 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
0 0 0 0 1 0 1 1 0 1 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 1 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1
1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0
```

Figure 4. Binary indicator sequence

The complex indicator sequence is defined by replacing the nucleotide with 1, -1, iota and -iota values.

Figure 5 shows a portion of gene AF099922 after application of complex indicator sequence.

```
 0+1i;0+1i;-1;1;1;0+1i;0+1i;0-1i;1;1;0+1i;0+1i;1;1;1;1;0-1i;1;0+1i;-1;0-1i
1i;1;1;1;1;1;1;0+1i;0+1i;0+1i;0+1i;0+1i;0+1i;0-1i;1;1;1;0-1i;0+1i;-1;-1;-1
1i;0+1i;0+1i;0+1i;-1;0-1i;0+1i;0-1i;0-1i;-1;1;1;0-1i;0+1i;0+1i;1;-1;0+1i;-
1;0+1i;-1;-1;-1;-1;0-1i;0+1i;1;1;1;0+1i;1;0+1i;0+1i;0+1i;0-1i;0+1i;1;-1;0+
1;1;1;-1;0+1i;0-1i;1;0-1i;1;0-1i;-1;-1;0-1i;0+1i;0-1i;0+1i;-1;-1;1;-1;0+1i
1i;0+1i;0+1i;0+1i;0+1i;0-1i;-1;0-1i;1;1;-1;0-1i;0-1i;1;1;1;1;1;1;1;0-1i;
1i;0-1i;-1;0-1i;0-1i;1;0-1i;1;0-1i;0+1i;0-1i;-1;-1;0-1i;1;0+1i;0+1i;0+1i;1
1i;0+1i;0+1i;0-1i;0-1i;0+1i;0+1i;-1;-1;1;1;0+1i;-1;0+1i;0-1i;0-1i;1;-1;0+1
1i;1;0+1i;0+1i;0+1i;0+1i;0-1i;1;1;-1;1;0+1i;0+1i;0+1i;0+1i;0+1i;0-1i;0-1i;
1i;1;1;1;1;1;0+1i;0+1i;1;0+1i;0+1i;0+1i;0+1i;0+1i;1;0-1i;0+1i;-1;0+1i;1;-1
1i;-1;0+1i;1;0-1i;0-1i;1;-1;0-1i;-1;-1;-1;1;0+1i;0+1i;0+1i;0+1i;0+1i;0+1i;
1i;1;1;0+1i;1;1;0+1i;0+1i;0+1i;0+1i;0+1i;0+1i;0+1i;-1;0-1i;0+1i;-1;0+1i;0+
1;1;0+1i;0+1i;1;-1;0+1i;0-1i;0-1i;-1;0-1i;1;1;0+1i;0+1i;0+1i;0+1i;-1;0-1i;
1i;0+1i;1;0-1i;-1;1;1;1;1;0+1i;0-1i;0+1i;1;-1;0+1i;0+1i;1;1;0+1i;0+1i;0+1i
1i;1;0+1i;1;0+1i;0+1i;0+1i;0+1i;0+1i;1;-1;0-1i;0-1i;0+1i;0-1i;0-1i;0+1i;0+
1;1;1;1;0-1i;1;1;1;1;1;1;0+1i;0+1i;-1;0+1i;-1;0+1i;0+1i;0-1i;1;0-1i;0+1i;0
+1i;0+1i;0+1i;0+1i;0-1i;1;0-1i;1;0+1i;0+1i;1;0+1i;0+1i;0+1i;0+1i;1;0+1i;-1
1i;1;0+1i;0+1i;0+1i;0+1i;0-1i;0+1i;-1;-1;1;1;0+1i;0+1i;0+1i;0-1i;0-1i;0-1i
1i;-1;1;-1;-1;-1;0+1i;0+1i;1;0-1i;0+1i;-1;0+1i;1;-1;1;1;-1;0+1i;1;0-1i;1;0
1i;0-1i;-1;0-1i;-1;-1;-1;1;0+1i;0+1i;0+1i;0+1i;0+1i;0+1i;-1;1;0+1i;0+1i;0+
1i;0+1i;0+1i;0+1i;0+1i;-1;0-1i;-1;-1;0+1i;0+1i;-1;0+1i;-1;0+1i;0-1i;0-1i;0
```
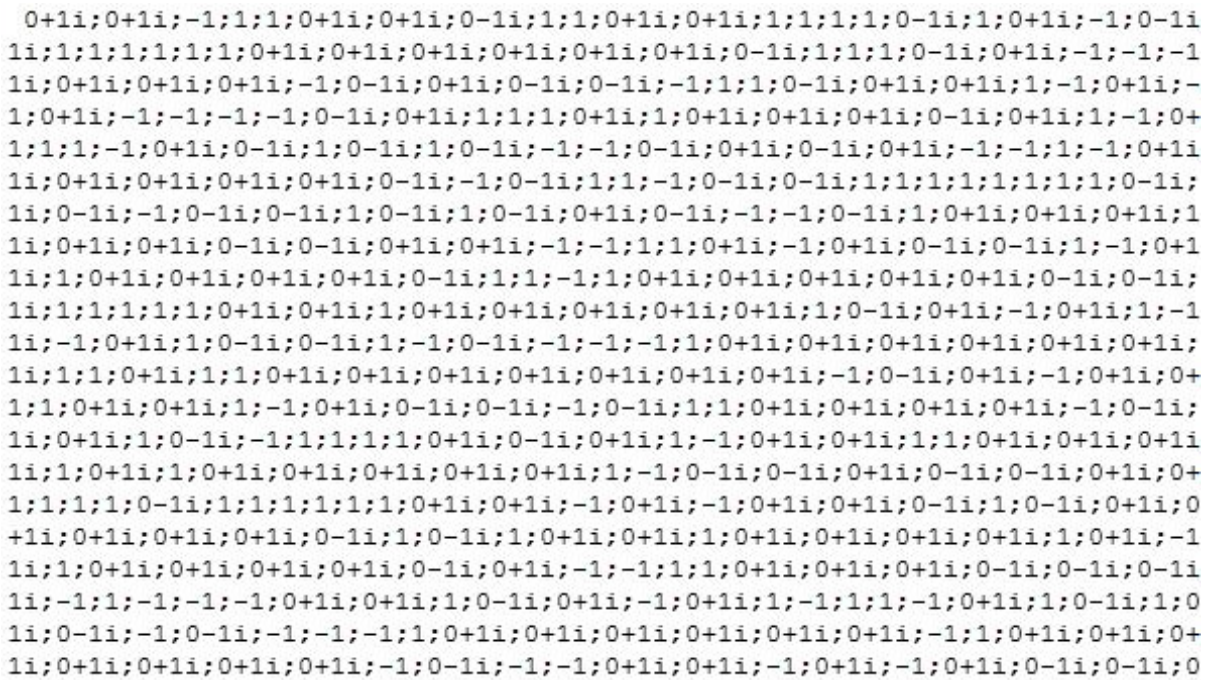
Figure 5. Complex indicator sequence applied to gene

The complex indicator sequence transforms the sequence into four digital patterns with associated weights. It is worth mentioning that this indicator sequence provided close range estimation for nucleotides in the literature.

This signal is then passed through the steps of windowed STFT for exonic prediction spectral analysis. This helps to extend the length of the signal to a target length so that perfect analysis could be performed over the signal.

Figure 6 shows that signal has been extended to a desired length. The length of signal was 8000 patterns. The convolution method suggests that to perform a better approximation, the signal should be extended to 8192 patterns. The signal should be mapped employing Kaiser Window of length 351 base pairs. The previous power of two shows a numerical value 4096 which truncates the signal from its original length. Truncation phenomenon can degrade the results and may bring faulty approximation that would lead to differ from the standard range of exons.

Figure 7 depicts the wavelet sketch for db3 wavelet. Scaling and wavelet functions have been described. Decomposition of low pass filter and high pass filters have been identified, similarly signal synthesis for low and high pass filters have been shown. This sketch demonstrates that signal should be passed through these defined filters to further analyze it for denoising and enhancement. The upward and downward curves self explains the convolution of signal with the window function at desired location of nucleotides.
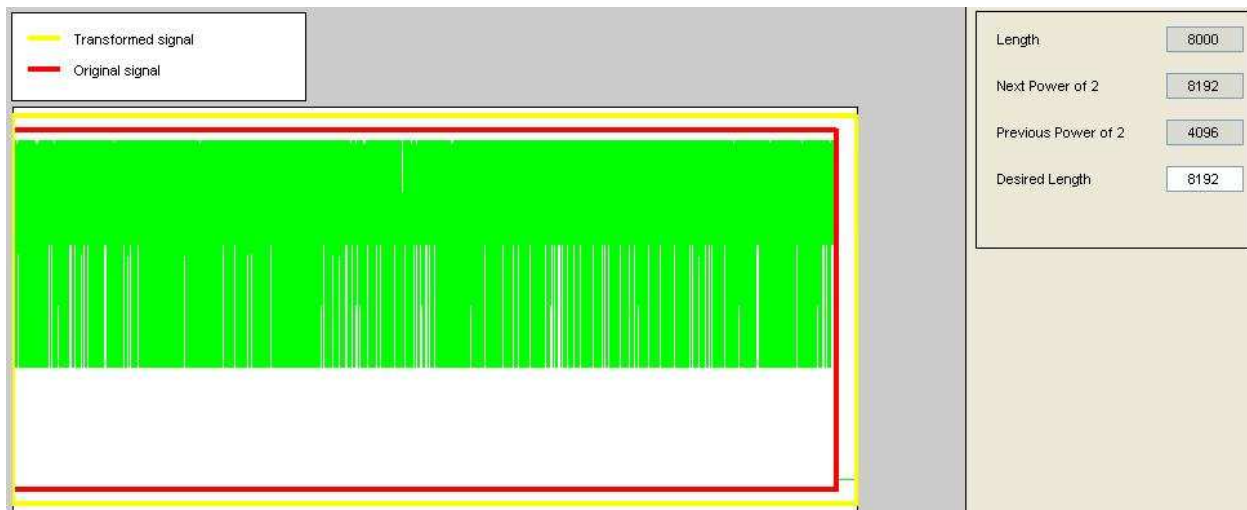
Figure 6. Signal extension to desired length
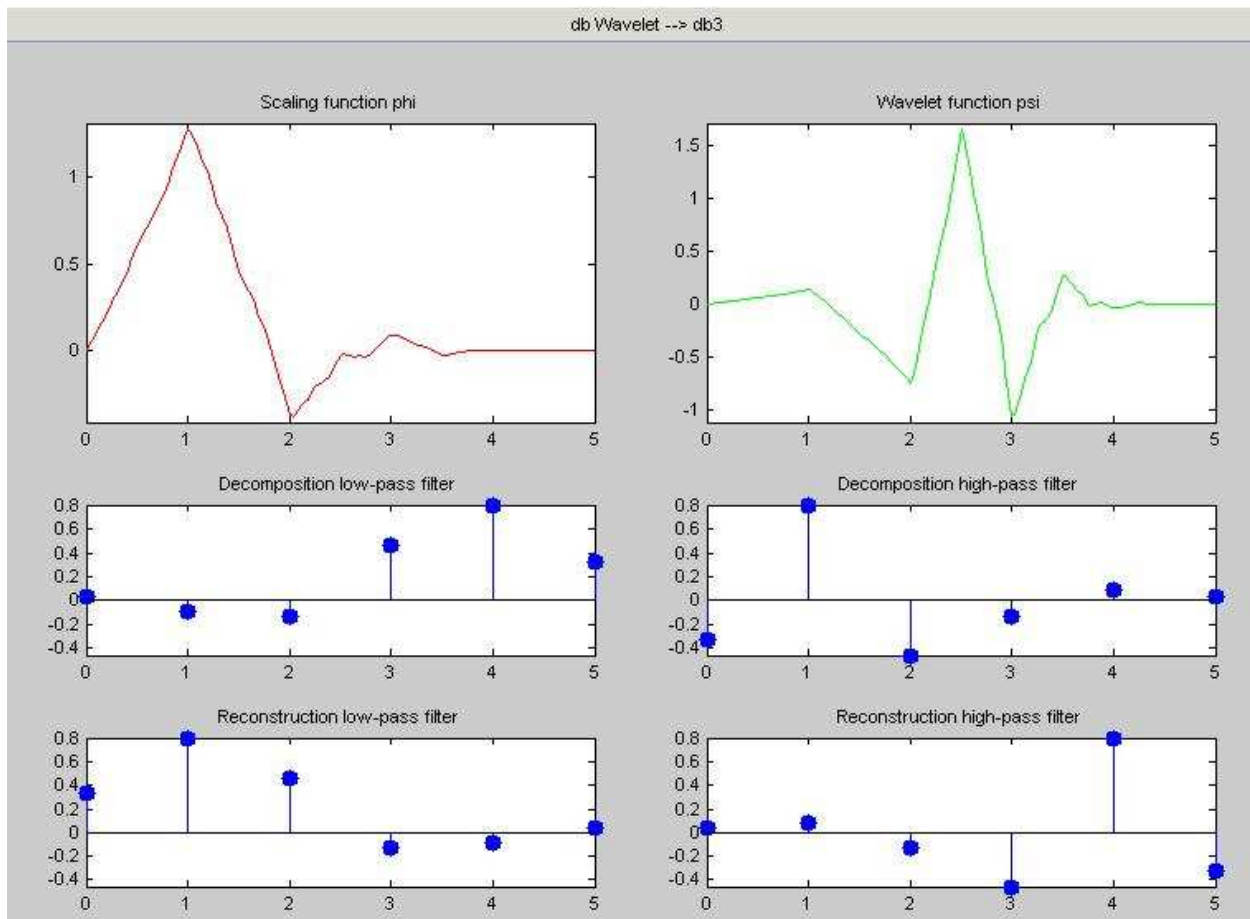
Figure 7. Wavelet of db3 sketch

Figure 8 shows a wavelet tree for Shannon entropy of signal. There is a tree structure for the nodes depicting different position factors. Colored coefficients for terminal nodes can be observed. The first rectangle at the right top shows the analyzed signal at different nucleotides places (diffusion of bases at DNA strand). Calculation of Shannon entropy would assist in further identification of boundary values for individual nucleotides at power spectral density estimation graphs.

The digital signal passes through refinement stages. First, the sequence was obtained as a raw data which was purified to access only nucleotides bases without degrading factor. This is termed as an important process because any kind of unwanted characters may lead to different set of nucleotides values that would be away from actual results.

The digital signal under discussions contains 8000 base pairs. The same dataset was used extensively in literature by other researchers and it is being used as a bench mark. The spectral estimation graph reveals that it contains five exonic regions at different nucleotides ranges. Identification of these ranges close to standard range demands to denoise the signal and selection of an appropriate window function that could be used for perfect convolution. The standard convolution requires to multiply the signal with a portion of window function, this is the reason that signal was extended to a power of two to make it to desired length. Each

frame of the signal is calculated numerically equal sized so that power spectral graph is uniform in all characteristics.

For discrete wavelet transforms of order three, the signal is decomposed and synthesized. These db3 performs the quick vanishing of coefficients for approximate and detail patterns.

Figure 9 shows a glimpse of original signal. There are 8000 base pairs shown in the form of a digital pattern. Cumulative histogram of signal shows different range of weight values assigned to nucleotides base pairs. It can be seen that nucleotides with numerals higher than 0.25 have high frequency while those between 0.1 and less than 0.25 have lower frequency. The individual histogram also shows three separate characterizations of nucleotide weight values. The standard deviation has been found to be 0.09037, median of absolute deviation is 0.11 while mean absolute deviation is 0.07843. The maximum range is 0.375 while minimum range is 0.125 and the average range is depicted as 0.25.
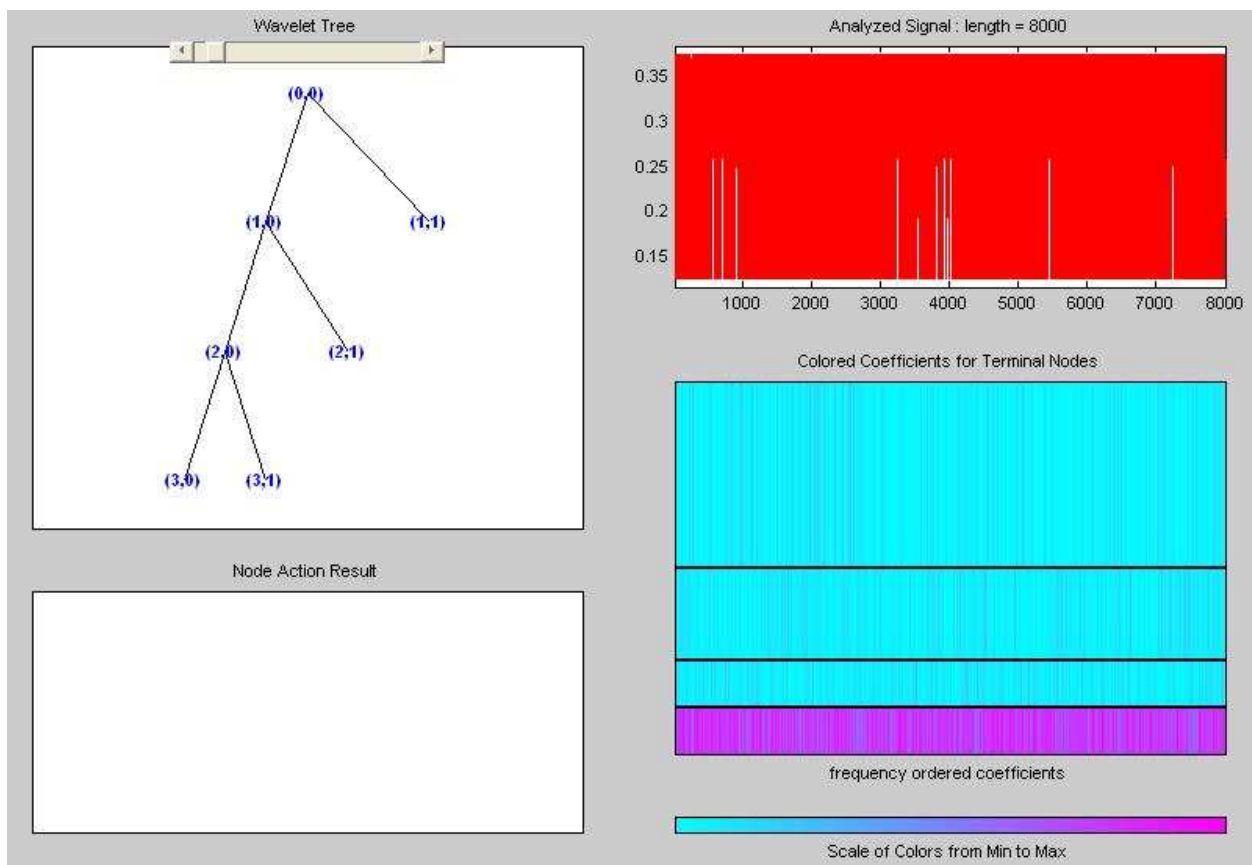

Figure 8. Wavelet tree for Shannon entropy
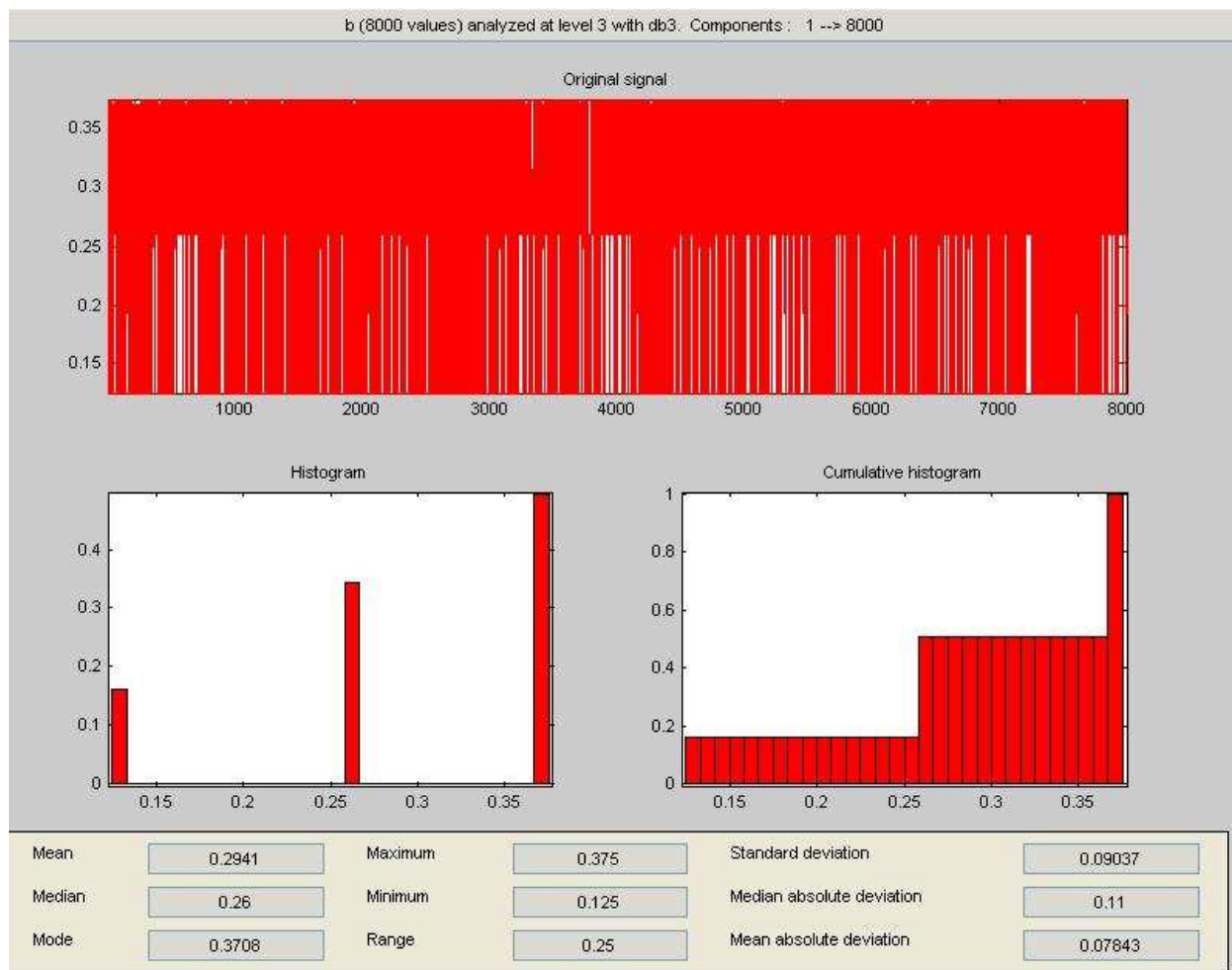
Figure 9. Original signal at level 3

The histogram calculates the frequency of nucleotides bases in the signal. Since the signal was mapped with an enhanced indicator sequence which assigns perfect weights to nucleotides bases, histogram of such a signal is uniform. It is observed that almost half of the signal is diffused in first band and half in the second band. First half band shows the smaller histogram values (frequency components) while second half band depicts some larger histogram values.

The individual histogram components dependant over the individual nucleotide bases, for instance, the numeric value of Adenine is 0.260, which is plotted against the other numeric values for bases in individual histogram. Depending over the weights assigned to Thymine and cytosine, the histogram shape may change.

It is also important to note that histogram of frequency components present the redundancy of bases in the digital pattern. This repetition depends over the order of nucleotides in DNA sequence which defines the habits, traits and other characteristics of species.

Figure 10 shows the synthesized signal at level three. Like the original signal, the synthesized signal owns the same histogram characteristics. There are 8000 base pairs shown in the

form of a digital pattern. Cumulative histogram of signal shows different range of weight values assigned to nucleotides base pairs. It can be seen that nucleotides with numerals higher than 0.25 have high frequency while those between 0.1 and less than 0.25 have lower frequency. The individual histogram also shows three separate characterizations of nucleotide weight values. The standard deviation has been found to be 0.09037, median of absolute deviation is 0.11 while mean absolute deviation is 0.07843. The maximum range is 0.375 while minimum range is 0.125 and the average range is depicted as 0.25.
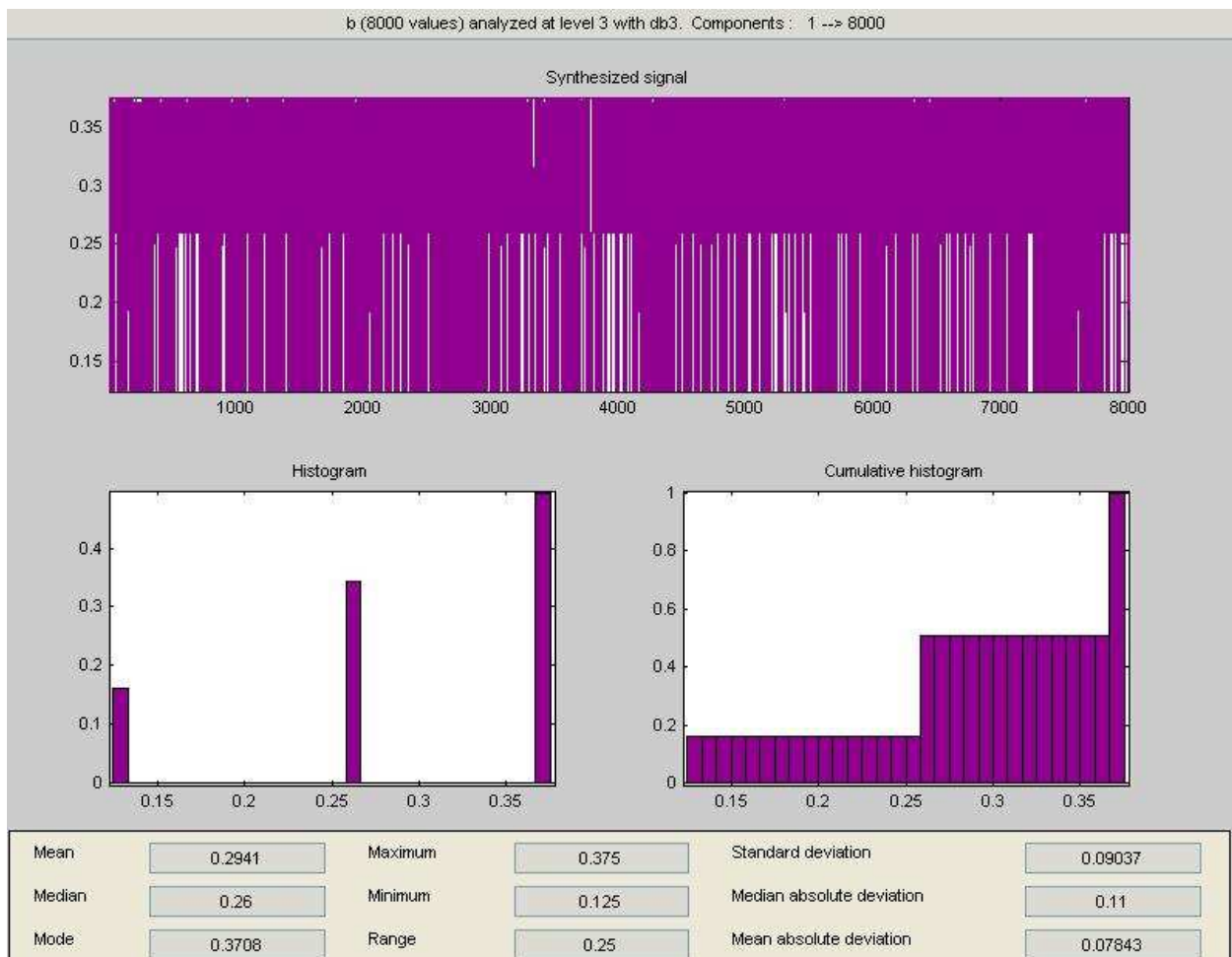
Figure 10. Synthesized signal at level 3

The synthesized signal shows the same histogram even after its decomposition. The synthesized signal is perfectly reconstructed by employing discrete wavelet transforms. The approximate and detail coefficients of signal are obtained in passing through a series of filters. These digital filters have been defined and constructed using Matlab. The decomposed signal is addition of approximate and detail coefficients at level three along with detail coefficients at level two and level one.

As for as, we decompose the signal, the components are loosely packed.

Figure 11 depicts the signal decomposition into approximate and detail coefficients. Symbol s represents the original signal. Approximate and detail coefficient at level three show the reduced complexity in the signal.
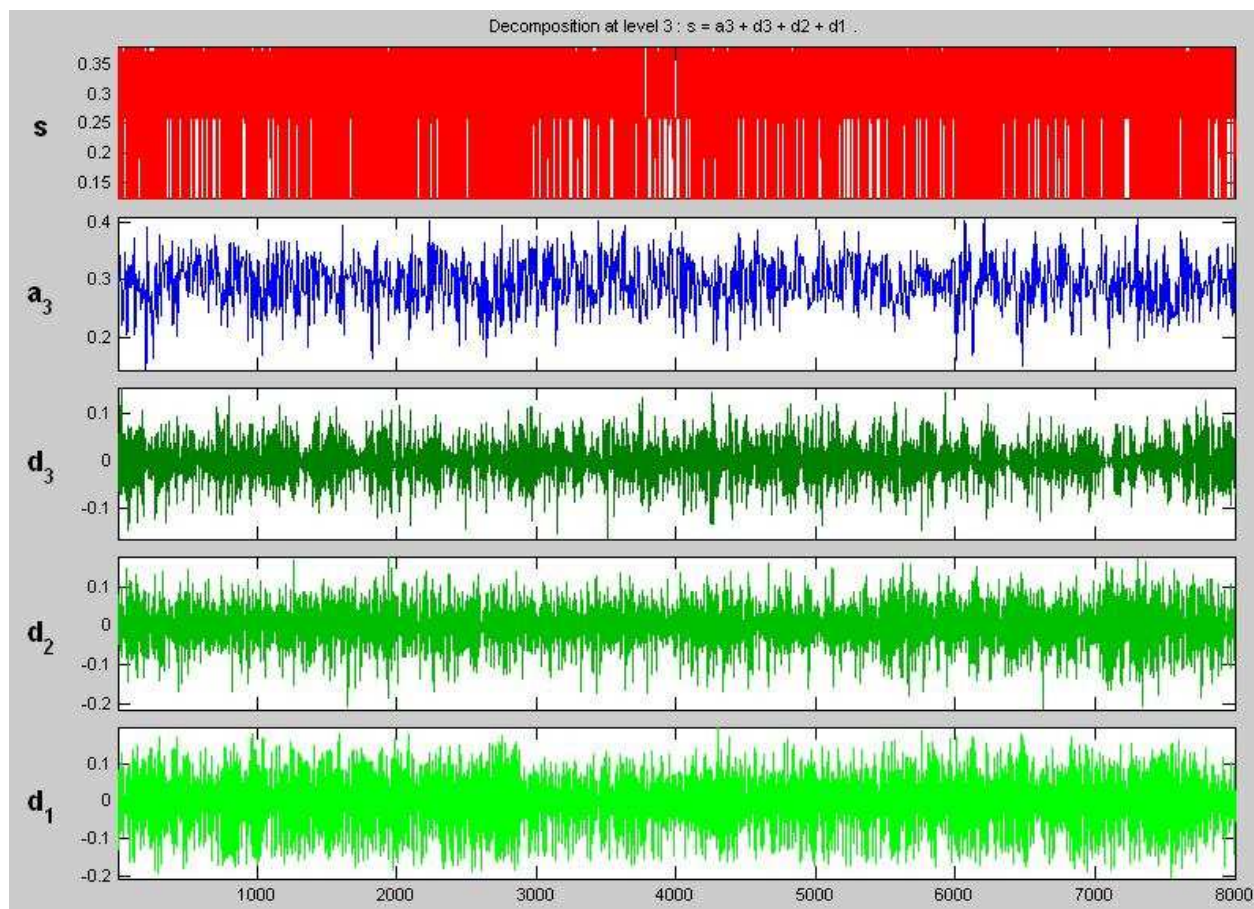
Figure 11. Signal decomposition

Figure 12 presents the histogram for approximate and detail coefficients. At level one, the concentration of components is less than other levels. Level two shows that signal components are more concentrated. At level three, the signal components are more closely packed. Likewise, the histogram for approximate coefficients presents the same phenomenon. At level one, the concentration of components is less than other levels. Level two shows that signal components are more concentrated. At level three, the signal components are more closely packed. It can be observed that original signal and synthesized signal contain the same number of components. The concentrations of signal components are uniform over these histograms' plots, which depict the perfect reconstruction of digital signal.
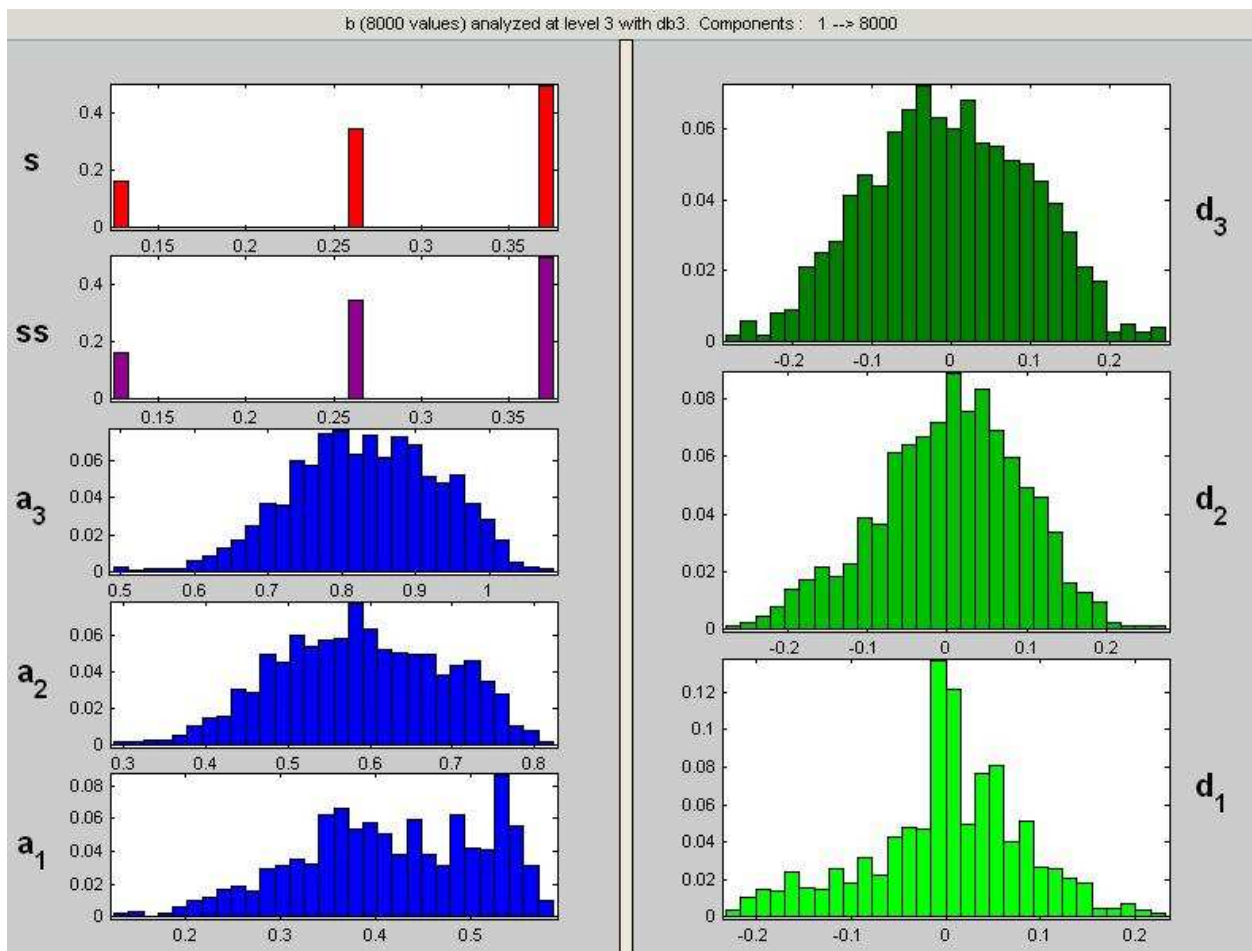
Figure 12. Histogram of signal

Figure 13 shows the density estimation of approximate and detail coefficients. The density estimate of original signal shows the numerals for nucleotides present in the signal in digital format as a general. The approximate coefficients at level three presents a sharp peak at some 0.25 points. The signal remains uniform through the course except at another peak value ranging from 0.37 to the end of the signal. The density estimation for detail coefficients at level one shows the same sharp peak around 0.27 points. The same peak can be observed around 0.40 at level two. At level three, the phenomenon is same but the signal components are loosely packed than level two. At granular level, the components are more packed at level one than other levels.
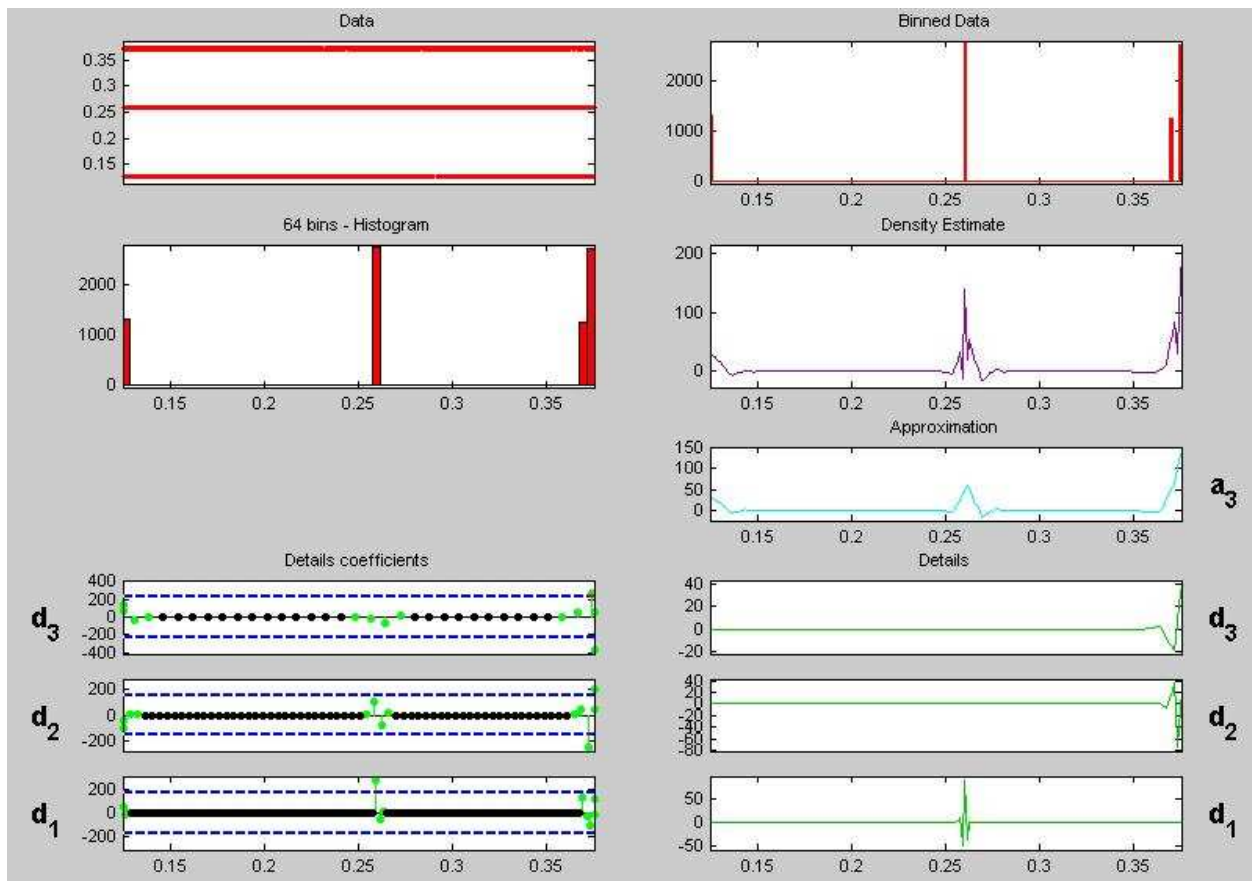
Figure 13. Density estimation of signal

Figure 14 shows the resultant denoised signal. It is obvious that preview of detail coeffi‐
cients at level three shows the loosely packed signal components. The original signal is rep‐
resented in red color. The threshold coefficients are shown in vertical bars for all nucleotide
range (8000 base pairs). The coefficients at detail level depicts a hierarchy of packed, loosely
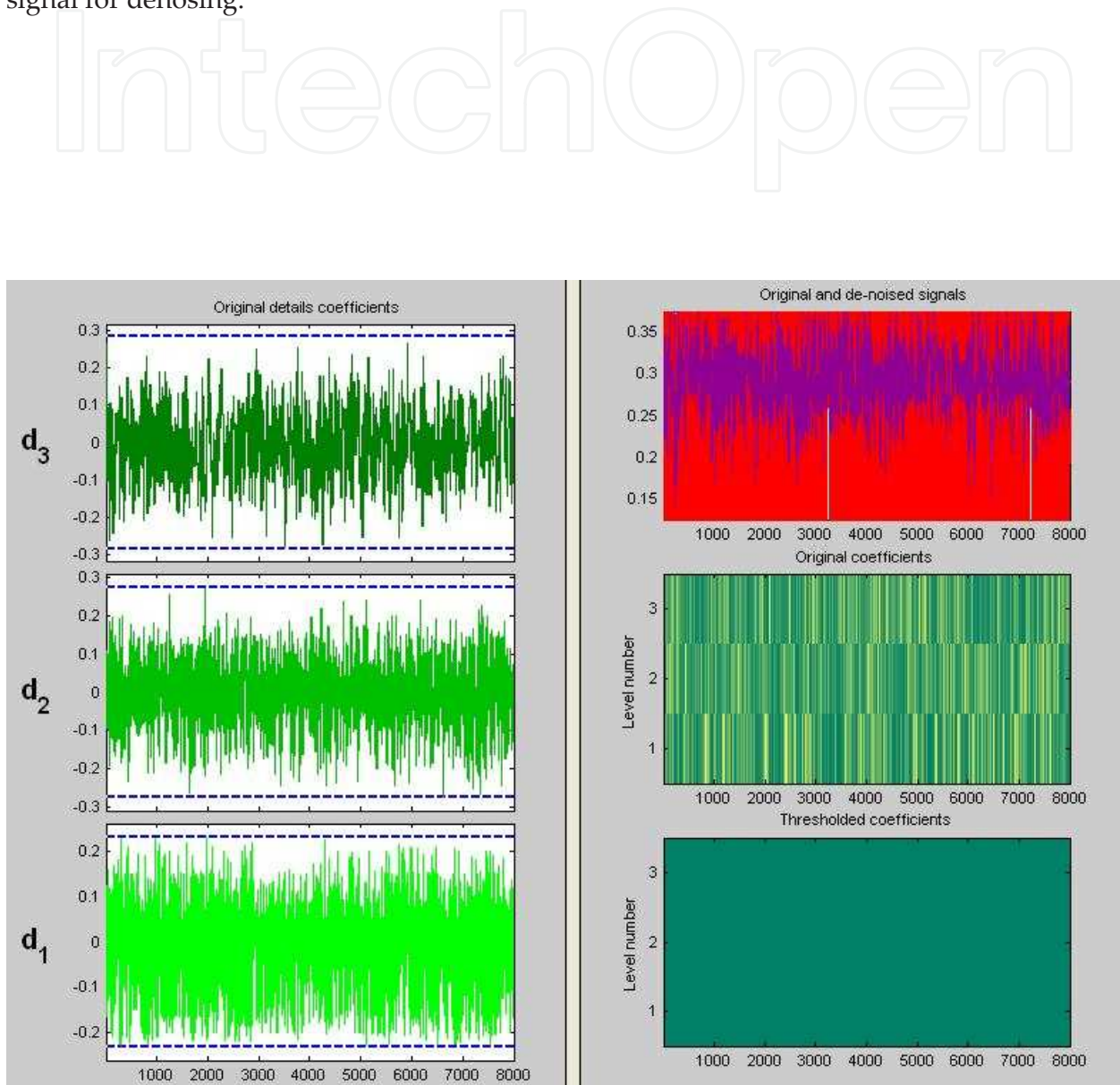packed and more loosely packed components, which shows a gradual improvement in the
signal for denosing.

Figure 14. Denoised signal

Figure 15 shows the approximate coefficients at level three. A sharp gradual change can be observed in a commutative histogram. The peaks are more pronounced at from point one to onwards. In another histogram, the peaks are not much visible around first 0.6 points, there is a sharp gradual increment in the bars reaching the maximum of 0.07 points then a gradual decrement is observed leading it to point one. The peaks are less pronounced after this point. The coefficients of approximation at level three show the signal as loosely packed components.

It can be observed that detail coefficient at level one are packed showing more concentration of nucleotides while detail coefficients at level two are loosely packed. The coefficients at level three are more significant than other levels, which represents that the signal is filtered for refinement. The signal was passed through a series of filters for the wavelet db3 which denoised the signal as a result of reconstruction of signal.
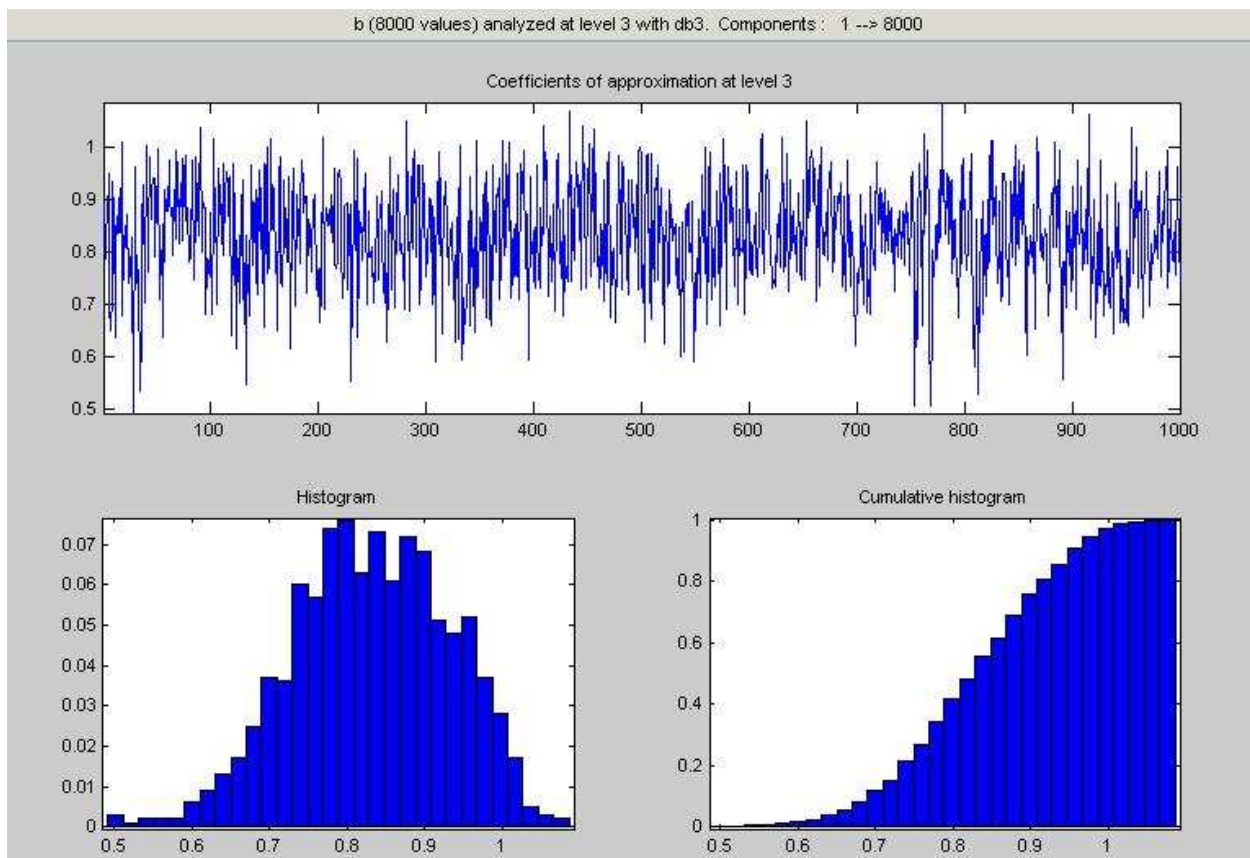

Figure 15. Approximation coefficients at level 3

Figure 16 shows the detail coefficients at level three. A sharp gradual change can be observed in a commutative histogram. The peaks are more pronounced at from point one to onwards. In another histogram, the peaks are not much visible around first 0.6 points, there is a sharp gradual increment in the bars reaching the maximum of 0.07 points then a gradual decrement is observed leading it to point one. The peaks are less pronounced after this point. The coefficients of detail at level three show the signal as loosely packed components.
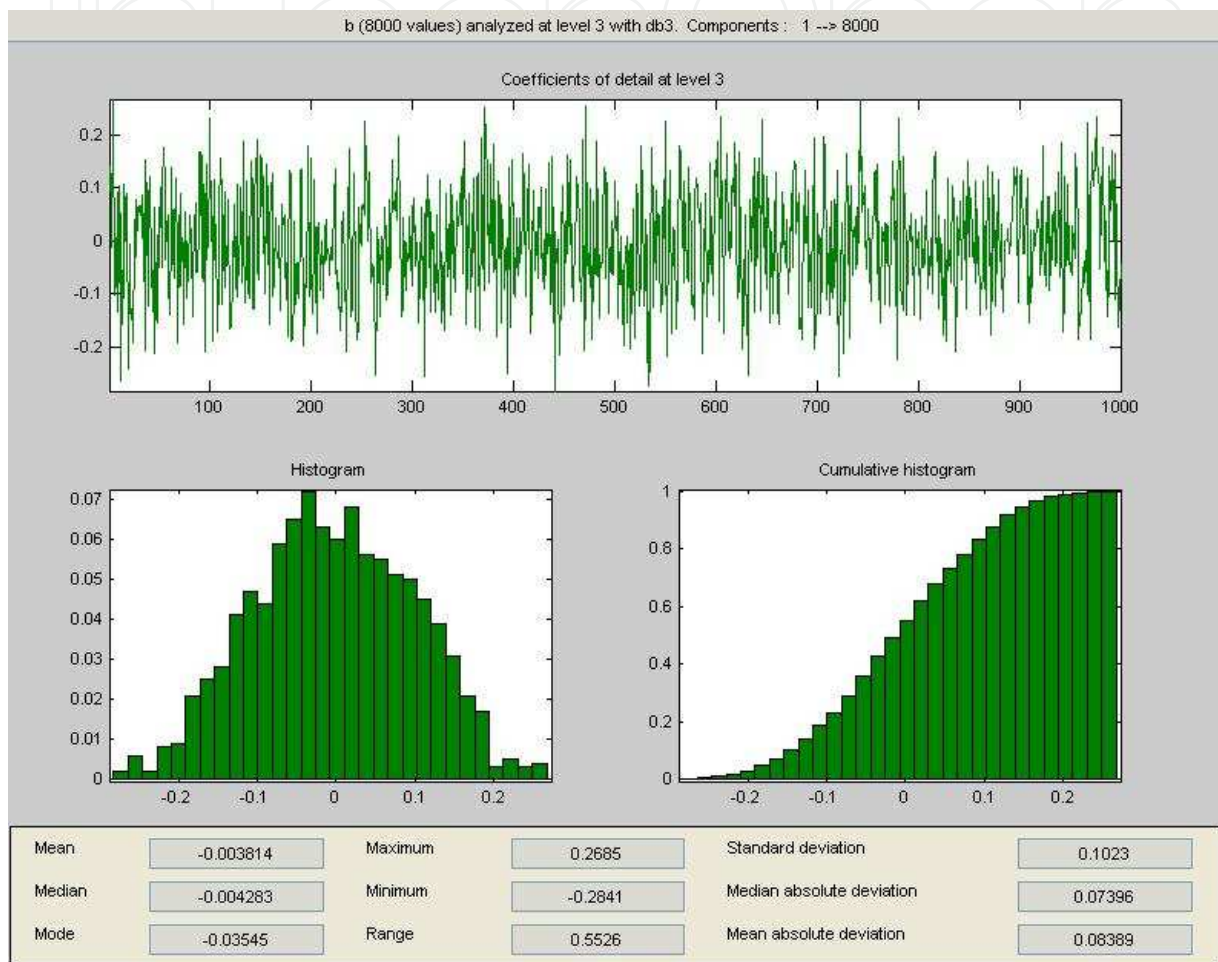

Figure 16. Coefficients of detail at level 3

It can be observed that detail coefficient at level one are packed showing more concentration of nucleotides while detail coefficients at level two are loosely packed. The coefficients at level three are more significant than other levels, which represents that the signal is filtered for refinement. The signal was passed through a series of filters for the wavelet db3 which denoised the signal as a result of reconstruction of signal.

Table 1 presents the nucleotide range for exons. Clear differences can be observed as a comparative analysis of various approaches. Binary and EIIP methods show a wide range difference compared with the standard range. Complex method results are better than the first two approaches. Digital filters behave accordingly. The proposed approach has more significant results than other prevailing approaches.

Method	E_1	E_2	E_3	E_4	E_5
Binary Method	656-1206	2406-3106	3806-4406	5306-5806	7106-7706
EIIP Method	706-1206	2206-2906	3906-4406	5206-5806	7206-7706
Complex Method	750-1100	2600-2906	3600-4406	5206-5706	7106-7600
Filter 1 (Anti-notch)	656-1206	2450-3106	3806-4450	5306-5850	7106-7750
Filter 2 (Multistage)	706-1250	2206-2950	3906-4450	5206-5850	7206-7706
proposed Method	750-1050	2450-2906	3950-4380	5206-5600	7220-7680
NCBI Range	928-1039	2528-2857	4114-4377	5465-5644	7255-7605

Table 1. Range of exons for different methods

4. Conclusion

Bioinformatics is a very rapidly emerging field of research. The genome sequence analysis is an interesting and challenging task that needs great attention. The analysis brings very promising relevance between species. The proposed approach provides a way to better identify the genetic regions in mixture of exon-intron noise. The focus directed to minimize the leakage of frequency contents by adoption of an optimal indicator sequence. We also reduced the signal noise by using Kaiser Window function with length 351 base pairs. The spectral density estimation was enhanced with application of wavelet transforms. The proposed dimensions reduced the noise and increased the sharp peaks of exons in density graphs. We have observed significant improvement in results as a comparative analysis between existing techniques and compared the results with strands NCBI range.

Author details

Noor Zaman[1*], Ahmed Muneer[1] and Fausto Pedro García Márquez[2]

*Address all correspondence to: nzaman@kfu.edu.sa

*Address all correspondence to: mmalik@kfu.edu.sa

*Address all correspondence to: FaustoPedro.Garcia@uclm.es

1 College of Computer Sciences & Information Technology,King Faisal University, Saudi Arabia

2 ETSI Industriales, Universidad Castilla-La Mancha, Ciudad Real, Spain

References

[1] Tina P George and Tessamma Thomas, "Discrete wavelet transform de-noising in eukaryotic gene splicing", BMC Bioinformatics, 11(Suppl 1):S50, doi: 10.1186/1471-2105-11-S1-S50, 2010

[2] Roy, M., Biswas, S. and Barman, S., "Identification and Analysis of Coding and Noncoding Regions of a DNA Sequence by Positional Frequency Distribution of Nucleotides (PFDN) Algorithm", 4th International Conference on Computers and Devices for Communication. CODEC 2009, Page(s): 1 - 4, 2009

[3] GuoShuo and Zhu Yi-sheng, "Prediction of Protein Coding Regions by Support Vector Machine", International Symposium on Intelligent Ubiquitous Computing and Education, Digital Object Identifier: 10.1109/IUCE.2009.141, Page(s): 185 - 188, 2009

[4] J. Quintanilla-Domínguez, B. Ojeda-Magaña, J. Seijas, A. Vega-Corona, D. Andina, "Edges Detection of Clusters of Microcalcifications with SOM and Coordinate Logic Filters", Proceedings of the 10th International Work-Conference on Artificial Neural Networks, Pages: 1029 - 1036, ISBN:978-3-642-02477-1, 2009

[5] Ruchira. Ajay Jadhav, Roopa Ashok Thorat, "Computer aided breast cancer analysis and detection using statistical features and neural networks", Proceedings of the International Conference on Advances in Computing, Communication and Control, Mumbai, India, ISBN:978-1-60558-351-8, 2009

[6] Muneer Ahmad and Hassan Mathkour, "Multiple Sequence Alignment with GAP consideration by pattern matching technique", International Conference on signal acquisition and processing, Malaysia, 2009

[7] Muneer Ahmad and Hassan Mathkour, "Genome sequence analysis by Matlab histogram comparison between Image-sets of genetic data", WCSET, Hong Kong, 2009

[8] Muneer Ahmad and Hassan Mathkour, "A pattern matching approach for redundancy detection in bi-lingual and mono-lingual Corpora", IAENG 2009

[9] HazrinaYusofHamdani and SitiRohkmahMohdShukri, "Gene prediction system", International Symposium on Information Technology, Volume: 2, Digital Object Identifier: 10.1109/ITSIM.2008.4631728, Page(s): 1 - 7, 2008

[10] ShuoGuo and Yi-Sheng Zhu, "An integrative algorithm for predicting protein coding regions", IEEE Asia Pacific Conference on Circuits and Systems,Digital Object Identifier: 10.1109/APCCAS.2008.4746054, Page(s): 438 - 441, 2008

[11] Kakumani, R., Devabhaktuni, V., and Ahmad, M.O., "Prediction of protein-coding regions in DNA sequences using a model-based approach", IEEE International Symposium on Circuits and Systems, Digital Object Identifier: 10.1109/ISCAS.2008.4541818, Page(s): 1918 - 1921, 2008

[12] Akhtar, M., Ambikairajah, E. and Epps, J., "Optimizing period-3 methods for eukaryotic gene prediction", IEEE International Conference on Acoustics, Speech and Signal

Processing, Digital Object Identifier: 10.1109/ICASSP.2008.4517686, Page(s): 621 - 624, 2008

[13] Hota, M.K. and Srivastava, V.K., "DSP technique for gene and exon prediction taking complex indicator sequence", IEEE Region 10 Conference, Digital Object Identifier: 10.1109/TENCON.2008.4766667, Page(s): 1 - 6, 2008

[14] Dosay-Akbulut, M., 2006. Group I introns and splicing mechanism and their present possibilities in elasmobranches. J. Boil. Sci., 6: 921-925. DOI: 10.3923/jbs.2006.921.925

[15] Parent, A., I. Benzaghou, I. Bougie and M. Bisaillon, 2004. Transcription and mRNA processing events: The importance of coordination. J. Biological Sci., 4: 624-627. DOI: 10.3923/jbs.2004.624.627

[16] S. Al Wadi, MohdTahir Ismail, M.H. Alkhahazaleh and SamsulAriffinAddulKarim, "Orthogonal Wavelet Transforms in Forecasting Volatility: An Expermintal Results", World Applied Sciences Journal, Volume 10 number 3, 2010

[17] M. Hashemi, S. Khanzadi and A. Jamshidi, " Identification of Escherichia coli O157:H7 Isolated from Cattle Carcasses in Mashhad Abattoir by Multiplex PCR", World Applied Sciences Journal, Volume 10 number 6, 2010

[18] A. Ali, F. Khorshid, H. Abu-araki and A.M. Osman, ,"Tumor Lung Cancer Model for Assessing Anti-neoplastic Effect of PMF in Rodents: Histopathological Study", Trends in Applied Sciences Research Volume 6, Number 10, 1214-1221, 2011

[19] J. Singh, S. Gupta and H. Kaur , "Prediction of in vitro Drug Release Mechanisms from Extended Release Matrix Tablets using SSR/R2 Technique", Trends in Applied Sciences Research Volume 6, Number 4, 400-409, 2011